doi:10.1093/nar/gki061

HCAD, closing the gap between breakpoints and genes

Robert Hoffmann*, Joaquin Dopazo[1], Juan C. Cigudosa[1] and Alfonso Valencia

National Center of Biotechnology, CNB-CSIC, Campus de la UAM, Cantoblanco, Madrid 28049, Spain and
[1]Centro Nacional Investigaciones Oncologicas (CNIO), Melchor Fernández Almagro, 3, Madrid 28029, Spain

Received July 13, 2004; Revised and Accepted October 4, 2004

ABSTRACT

Recurrent chromosome aberrations are an important resource when associating human pathologies to specific genes. However, for technical reasons a large number of chromosome breakpoints are defined only at the level of cytobands and many of the genes involved remain unidentified. We developed a web-based information system that mines the scientific literature and generates textual and comprehensive information on all human breakpoints. We show that the statistical analysis of this textual information and its combination with genomic data can identify genes directly involved in DNA rearrangements. The Human Chromosome Aberration Database (HCAD) is publicly accessible at http://www.pdg.cnb.uam.es/UniPub/HCAD/.

In model systems, identifying and generating mutations is the usual genetic approach to understanding the function of individual genes. In humans, natural mutations, such as chromosome aberrations, are a comparable resource for genetic research, since DNA breakage and reciprocal recombination often lead to the fusion or deregulation of genes (1–3). Indeed, most human cancers (both leukaemias and solid tumours) and congenital disorders (including dysmorphology syndromes) display recurrent chromosome abnormalities. So far, the Mitelman database (http://cgap.nci.nih.gov/Chromosomes/Mitelman) constitutes the main effort to collect clinical and morphological data on cancer related chromosome aberrations (4). However, the Mitelman database contains relatively little molecular information and depends completely on manual curation.

We employ automatic text mining methods on PubMed abstracts to gather molecular and clinical facts for all recurrent and non-recurrent breakpoints described in human disorders. The complete and automatic identification of relevant abstracts from 10 million publications is possible, since aberration codes, for instance t(9;22)(q34;q11.2), are unambiguous. To facilitate the literature exploration of specific breakpoints, relevant information is compiled into textual

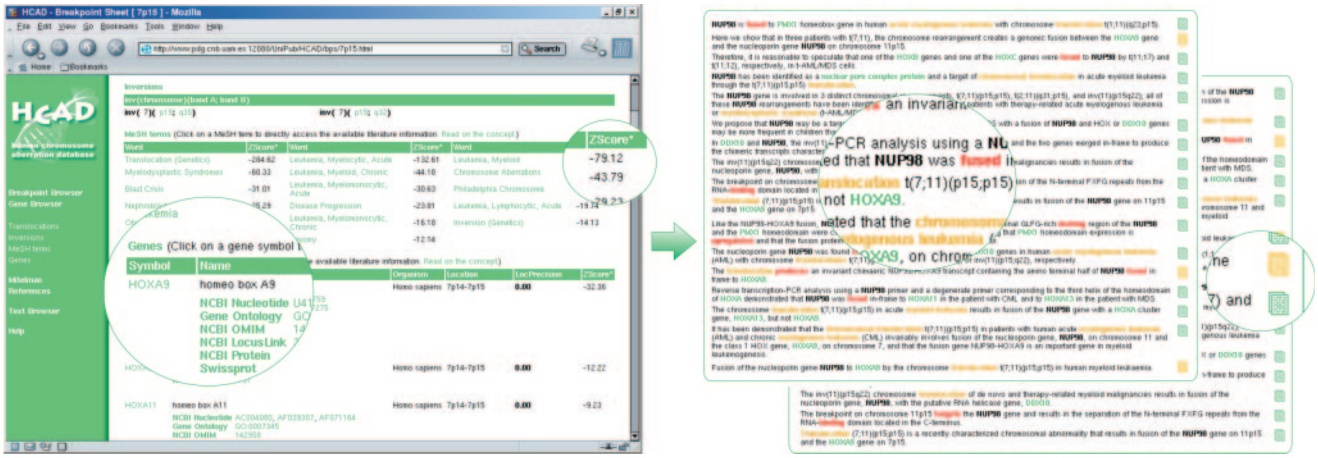

Figure 1. In the web interface, genes and biomedical terms serve as hyperlinks between information-rich sentences; iHOP concept, Information Hyperlinked over Proteins (7). In this manner, the information for breakpoints becomes accessible as a navigable network, allowing for a quick and intuitive exploration of the information for individual breakpoints and their associated genes and pathologies.

*To whom correspondence should be addressed. Tel: +34 91 5854570; Fax: +34 91 5854506; Email: hoffmann@cnb.uam.es
Correspondence may also be addressed to Alfonso Valencia. Tel: +34 91 5854570; Fax: +34 91 5854506; Email: valencia@cnb.uam.es

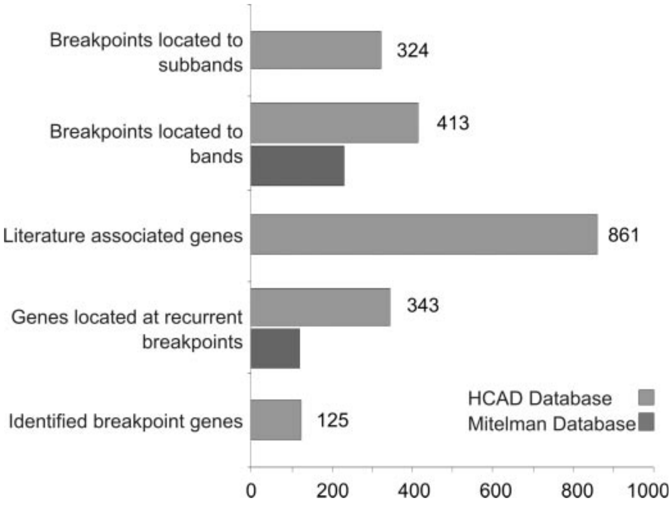

Figure 2. HCAD provides textual and comprehensive information for 737 breakpoints and integrates literature associations to 861 genes. The underlying identification of gene and protein names is based on a dictionary approach and obtains a precision of ~87%. However, since users can move directly between sentences from the source abstracts, they always retain the final say in determining the relevance and reliability of the retrieved information; this is an important improvement over existing (automatic and manual) information resources (4,5).

and comprehensive overviews (the database is updated every 1–3 months).

For every breakpoint, we calculated statistically significant genes (5,6) and biomedical terms (e.g. disease names) that were mapped back onto their source sentences. In the web-based interface these genes and keywords serve as hyperlinks between information-rich sentences (see Figure 1). In this manner, the information for breakpoints becomes accessible as a navigable network that is intuitive and exhibits all the advantages of the Internet (7). As researchers can move between sentences taken directly from source abstracts, they can always retain control over the reliability and significance of the information. Diseases and associative verbs are also highlighted and hyperlinked within the text to further facilitate the perception of associations with human pathologies. Additional molecular information on breakpoints is accessible through links to external databases, such as GenBank, LocusLink and OMIM. The complete system, called HCAD (Human Chromosome Aberration Database), contains 737 breakpoints and 861 *literature associated* genes from 2082 cytogenetically different translocations and inversions (see Figure 2).

The HCAD system was also designed to assist the identification of potential breakpoint genes. This is a difficult task

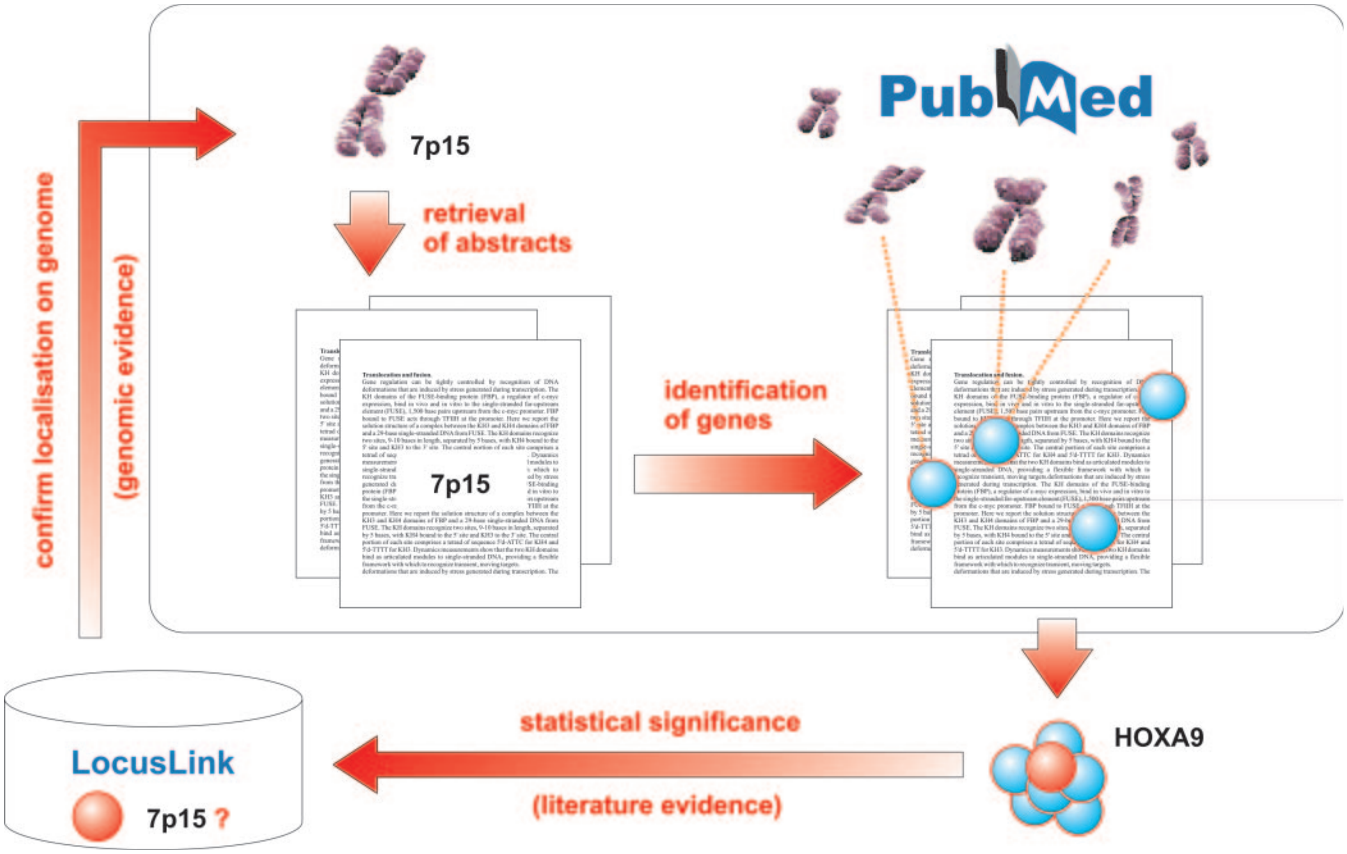

Figure 3. Identification of breakpoint candidate genes. Abstracts were retrieved from PubMed and clustered by breakpoints (e.g. 7p15). The genes found in a given cluster are not necessarily the actual breakpoint genes; however, the more often a gene is mentioned together with a breakpoint the more likely it is to be involved in an aberration (literature evidence). False positive associations are eliminated by crosschecking a gene's localization with genomic data (genomic evidence). The final decision on the relevance of a gene can then be confirmed by an expert through the web interface.

even though the complete human genome is now known, because of the sheer number of genes per cytoband (8). The premise behind HCAD is that genes directly affected by recurrent breakage events will be quoted more often in abstracts about the corresponding breakpoint, even if a direct proof for this association has not yet been described (see Figure 3). The statistical analysis in HCAD thus provides probabilities for genes to be relevant for a certain breakpoint (literature evidence). False positive associations of these predicted genes are eliminated by crosschecking their localization with genomic data (9). We found 343 of 861 *literature associated* genes to localize to recurrent breakpoints. Indeed, for one-third of these there are already clear experimental evidences that they are involved in fusion events. We believe that the HCAD information system provides a reliable basis for uncovering the role (10) of the remaining human genes in the context of chromosomal aberrations.

ACKNOWLEDGEMENTS

We thank Felix Mitelman for helpful discussion. We are grateful to the US National Library of Medicine for making MEDLINE publicly available. This work was supported in part by the ORIEL (IST-2001-32688) and TEMBLOR (QLRT-2001-00015) EC projects, MCYT projects BIO2001-0068 and SAF2001-0056, and the Fundacion Ramon Areces grant.

REFERENCES

1. Rabbitts,T.H. (1994) Chromosomal translocations in human cancer. *Nature*, **372**, 143–149.
2. Heim,S. and Mitelman,F. (1995) *Cancer Cytogenetics*, 2nd edn. Wiley-Liss, NY.
3. Vogelstein,B. and Kinzler,K.W. (2002) *The Genetic Basis of Human Cancer*, 2nd edn. McGraw-Hill Medical Publication Division, NY.
4. Mitelman,F., Johansson,B. and Mertens,F. (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genet*., **36**, 331–334.
5. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet*., **28**, 21–28.
6. Hoffmann,R. and Valencia,A. (2003) Life cycles of successful genes. *Trends Genet*., **19**, 79–81.
7. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nature Genet*., **36**, 664.
8. Rabbitts,T.H. (1999) Of methods and mapping. *Nature Med*., **5**, 24–25.
9. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet*., **16**, 44–47.
10. Tian,X.L., Kadaba,R., You,S.A., Liu,M., Timur,A.A., Yang,L., Chen,Q., Szafranski,P., Rao,S., Wu,L. *et al*. (2004) Identification of an angiogenic factor that when mutated causes susceptibility to Klippel–Trenaunay syndrome. *Nature*, **427**, 640–645.